Towards the Use of Relative Representations for Lower-Dimensional, Interpretable Model-to-Brain Mappings

T. Anderson Keller (takeller@fas.harvard.edu)
The Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University

Talia Konkle (talia_konkle@harvard.edu)
Harvard University, Psychology

Colin Conwell (cconwel2@jhu.edu)
Johns Hopkins University, Cognitive Science

Abstract

Current model-to-brain mappings are computed over thousands of features. These high-dimensional mappings are computationally expensive and often difficult to interpret, due in large part to the uncertainty surrounding the relationship between the inherent structures of the brain and model feature spaces. Relative representations are a recent innovation from the machine learning literature that allow one to translate a feature space into a new coordinate frame whose dimensions are defined by a few select 'anchor points' chosen directly from the original input embeddings themselves. In this work, we show that computing model-to-brain mappings over these new coordinate spaces yields brain-predictivity scores comparable to mappings computed over full feature spaces, but with far fewer dimensions. Furthermore, since these dimensions are effectively the similarity of known inputs to other known inputs, we can now better interpret the structure of our mappings with respect to these known inputs. Ultimately, we provide a proof-of-concept that demonstrates the flexibility and performance of these relative representations on a now-standard benchmark of high-level vision and firmly establishes them as a candidate model-to-brain mapping metric worthy of further exploration.

Keywords: Relative Representations, RDMs

Introduction

The mapping of artificial (deep) neural network (ANN) representations to biological brain activity is a now well-established method in cognitive computational neuroscience (Kriegeskorte, 2015; Yamins & DiCarlo, 2016). However, given the mismatch of dimensionality and inherent structure between ANN and brain representations, this mapping procedure is often complex and performed in a high dimensional space, limiting interpretability and inducing significant computational overhead.

In the machine learning literature, recent work by Moschella et al. (2023) has demonstrated that although different models (with different architectures, objectives and training data) may learn prima facie dissimilar representations in their original coordinate spaces, they surprisingly preserve the relative angles and distances between embedded points. In other words, the representations they learn are actually isometries of one another, and can be related by rotations, translations, reflections, and scaling. In order to discover the underlying common representation space, Moschella et al. (2023) suggested to instead embed points based on their cosine distance with respect to a fixed set of anchor points. At a high level, these 'relative representations' can be understood as extracting the unique fundamental similarity structure of the data which is being learned by these different models.

In this work, we extend this use of relative representations to the domain of model-to-brain mappings. In doing so we find a number of benefits and interesting directions for future research. Concretely, we find: (I) relative representations allow for high brain predictivity scores, comparable with the original representations, despite operating on a fraction of the dimensionality; (II) higher level visual regions (e.g. Occipital Temporal Cortex, OTC) appear to be a better match to these relative representations than lower level areas (e.g. Early Visual Cortex, EVC), potentially reflecting the fact that the abstract structure of the data extracted by relative representations is also better represented by OTC; and (III) relative representations for models can be seen as more interpretable since each axis now denotes similarity to a single known data-point 'anchor'.

Methods

In this section we describe how, given a stimulus set $\mathbb{X}$ (in our case images from the Natural Scenes Dataset of Allen et al. (2022)), a trained encoder model E, and corresponding functional Magnetic Resonance Imaging (fMRI) measurements, we can construct *relative representations* and use these to compute efficient and interpretable model-to-brain mappings.

Relative Representations Let $\mathbb{A}$ be a size N subset of the stimulus set $\mathbb{X}$. For each 'anchor' $\mathbf{a}^{(j)} \in \mathbb{A}$, compute its embedding as $\mathbf{e}_{\mathbf{a}^{(j)}} = E(\mathbf{a}^{(j)})$ for some encoder model E. Similarly, for each $\mathbf{x}^{(j)} \in \mathbb{X}$, compute its embedding as $\mathbf{e}_{\mathbf{x}^{(j)}} = E(\mathbf{x}^{(j)})$. The *relative representation* of $\mathbf{x}^{(j)}$ is then given as $\mathbf{r}^E_{\mathbf{x}^{(j)}} = \left(\mathrm{sim}\left(\mathbf{e}_{\mathbf{x}^{(j)}}, \mathbf{e}_{\mathbf{a}^{(1)}}\right), \ldots \mathrm{sim}\left(\mathbf{e}_{\mathbf{x}^{(j)}}, \mathbf{e}_{\mathbf{a}^{(N)}}\right)\right)$, for some similarity function sim. In this work, we use the cosine similarity function: $\mathrm{sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{xy}}{||\mathbf{x}||||\mathbf{y}||} = \cos(\theta_{x,y})$. For cosine similarity, the relative representations are invariant to global rotations, reflections, and rescalings, meaning if two encoders E and $\tilde{E}$ produce the same embeddings up to a rigid transformation, then $\mathbf{r}^E_{\mathbf{x}^{(j)}} = \mathbf{r}^{\tilde{E}}_{\mathbf{x}^{(j)}}$ is invariant to the choice of encoder.

Relative Representational Dissimilarity Matrix A common method for comparing representations from models and neural data involves a Representational Disimilarity Matrix (RDM) (Kriegeskorte, Mur & Bandettini, 2008). Explicitly, these matrices are computed by taking the Pearson's correlation coefficient (PCC) between the representations of all pairs of stimuli from a dataset, i.e. the i'th row and j'th column of the RDM is given as: $\mathrm{RDM}_{i,j} = \rho(\mathbf{e}_{\mathbf{x}^{(i)}}, \mathbf{e}_{\mathbf{x}^{(j)}}) \forall i, j$. Interestingly, we see that this is equivalent to a matrix of the aforementioned relative representations, where the anchor set is chosen to be equal to the full dataset, and the similarity function is chosen to be the PCC. The relative representations can therefore be seen as a select (significantly reduced) subset of rows of the RDM with a carefully chosen similarity function to induce desired invariances. To compute what we call a Relative RDM (RRDM), we compute relative representations for the model first, and then compute the RDM on this. Explicity: $\mathrm{RRDM}_{i,j} = \rho(\mathbf{r}^E_{\mathbf{x}^{(i)}}, \mathbf{r}^E_{\mathbf{x}^{(j)}}) \forall i, j$. In this work we only compute relative RDMs for the model embeddings in order to maintain the physical structure of neural recording sites in the fMRI data. In preliminary analysis, we find relative RDMs computed for fMRI data significantly reduce the similarity of model-to-brain mappings, and leave further analysis to future work.

Representational Similarity: cRSA, srprRSA & eRSA To get a scalar 'score' of a model-to-brain mapping, a 'representational similarity analysis' (RSA) procedure is undertaken using the model and brain RDMs. Following Conwell et al. (2023), we compute three canonical measures of similarity between the RDMs: Classical RSA (cRSA: computing the mean PCC between all elements of the two RDMs), Sparse Random Projection Ridge RSA (srprRSA: first apply sparse random projection to encodings, then compute ridge-regression from model to fMRI, measuring final correlation of regression), and Encoding RSA (eRSA: use ridge-regression above to compute a new RDM, and compare RDMs with mean PCC).

Models & Anchor Selection We compute each of the above measures for a representative suite of 60 models following Conwell et al. (2023). We first compute the baselines using the original representations, and then compute the 'relative' scores by using the model's Relative RDM formulation. In the present work we begin with a fixed set of randomly selected anchors for simplicity. We present comparisons for 10, 100, & 1000 randomly selected anchors. Note that this can yield orders of magnitude in dimensionality reduction of the model features since the final ANN features are usually 1000's of dimensions.

Interpretability In this work we posit that relative representations increase interpretabiltiy of model-to-brain mappings due to the fact that each dimension of $\mathbf{r}^E_{\mathbf{x}^{(j)}}$ now corresponds to a similarity with an exactly known input stimulus from the dataset. Although in this abstract we do not study this for space considerations, this property yields an inherent meaning to representation dimensions, allowing for future work to leverage these for more meaningful model-to-brain mappings.

Results

In Figure 1 we show a comparison of RSA scores computed on the original features (blue) and relative representations (orange) based on 100 anchors for the three RSA metrics in two regions (EVC (left) & OTC (right three)). We see that the scores based on relative representations are comparable to the original scores in all settings, and the gap is significantly reduced for OTC over EVC. Furthermore, we see that srprRSA and eRSA have equivalently small gaps between the original and relative embeddings. In Figure 2, we compare the impact of different numbers of anchors (10, 100, 1000) for cRSA on the OTC region. We see that while there is a small difference, they are surprisingly consistent even down to 10 anchor points, emphasizing the potential for dimensionality reduction.

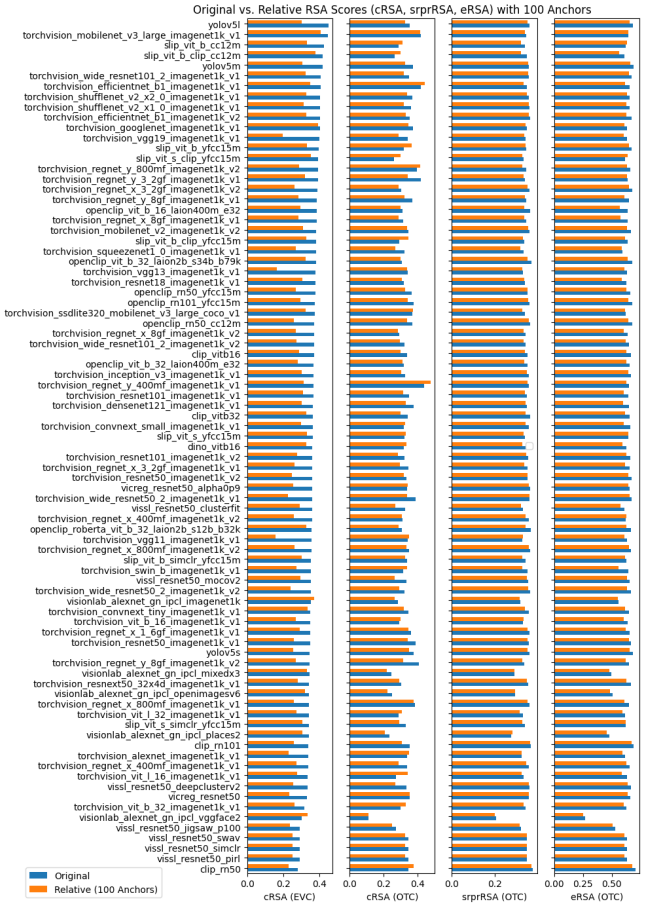

Figure 1: RSA scores for original and relative RDMs.

Summary

In summary, we find that by using relative representations we can remarkably reduce the dimensionality needed to represent model embeddings by up to two orders of magnitude while still achieving comparable RSA scores. We further find that these relative representations appear to match later visual areas more closely, warranting future research into their value as a candidate model-to-brain mapping metric. In future work it will be interesting to study more careful anchor selection procedures, leading to the hypothesized interpretability benefits and potentially even better RSA scores with fewer anchors.

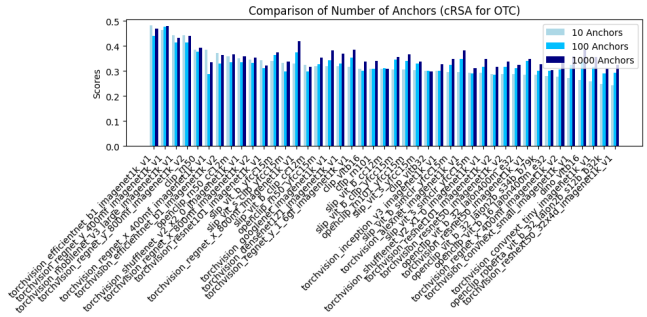

Figure 2: Comparison of cRSA with relative representations using differing numbers of anchors (10, 100, 1000) for OTC.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126. (Publisher: Nature Publishing Group US New York) doi: 10.1038/s41593-021-00962-x

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*. (Publisher: Cold Spring Harbor Laboratory) doi: 10.1101/2022.03.28.485868

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., & Rodolà, E. (2023). *Relative representations enable zero-shot latent space communication.*

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv preprint*. doi: 10.1101/407007

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356. (Publisher: Nature Publishing Group) doi: 10.1038/nn.4244